CRESST REPORT 803

EVALUATION OF SEEDS OF SCIENCE/ROOTS OF READING: EFFECTIVE TOOLS FOR DEVELOPING LITERACY THROUGH SCIENCE IN THE EARLY GRADES— UNIT ON PLANETS AND MOONS

JULY, 2011

Pete Goldschmidt
Hyekyung Jung

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Evaluation of Seeds of Science/Roots of Reading: Effective Tools for Developing Literacy through Science in the Early Grades—Unit on Planets and Moons

July, 2011

Pete Goldschmidt & Hyekyung Jung
CRESST/University of California, Los Angeles

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by grant number SA5415 from the SEEDS of Science/Roots of Reading Project with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

To cite from this report, please use the following as your APA reference:
Goldschmidt, P., & Jung, H. (2011). *Evaluation of Seeds of Science/Roots of Reading: Effective Tools for Developing Literacy through Science in the Early Grades—Unit on Planets and Moons* (CRESST Report 803). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

EVALUATION OF *SEEDS OF SCIENCE/ROOTS OF READING:EFFECTIVE TOOLS FOR DEVELOPING LITERACY THROUGH SCIENCE IN THE EARLY GRADES*—UNIT ON PLANETS AND MOONS[1]

Pete Goldschmidt & Hyekyung Jung
CRESST/University of California, Los Angeles

Abstract

This evaluation focuses on the Planets and Moons unit of the *Seeds of Science/Roots of Reading: Effective Tools for Developing Literacy through Science in the Early Grades* (*Seeds/Roots*) model of science-literacy integration. The evaluation is based on a cluster randomized design of 86 teachers, half of whom were in the treatment group. Multilevel models are employed to account for the clustering of students within teachers. Science outcomes are measured using science content, nature of science, and science inquiry outcomes, while literacy is measured using vocabulary, reading, and writing assessments. Additional analyses focus on the impact of teacher and student backgrounds, student attitudes, instructional methods, and teacher self-efficacy. Quantitative results indicate that the *Seeds/Roots* intervention demonstrates equivocal effects, statistically and substantively impacting student performance on the nature of science assessment and in vocabulary but not reading nor writing. There is suggestive evidence that the intervention improves science content performance. Teacher background and self-efficacy are generally unrelated to student performance, although teacher perception of implementation quality is related to outcomes. Although limited by data availability, exploratory results suggest that the intervention could be effective in English Language Learner (ELL-only) classrooms. Qualitative results indicate that treatment and control teachers noted that prior student literacy preparation impacted student performance, and empirical results, to some extent, corroborate the impact of this intake characteristic. Treatment teachers overwhelmingly found the *Seeds/Roots* unit usable, effective, and engaging—although requiring additional time to complete compared to the standard unit.

Introduction

This evaluation focuses on the *Seeds of Science/Roots of Reading: Effective Tools for Developing Literacy through Science in the Early Grades (Seeds/Roots)* model of science-literacy integration for Grade 5, developed and implemented by the Lawrence Hall of Science (LHS). The *Seeds/Roots* study is a multiyear project funded by the National Science Foundation. The project evaluation efforts build on previous *Seeds/Roots* evaluations (Goldschmidt & Jung, 2009; Wang & Herman, 2006) and focus on two major goals of the materials: usability and effectiveness. Formative evaluation processes (such as science

[1] We would like to acknowledge important contributions from the LHS staff who provided data and clarifications for the many inquiries we made.

assessment modification and rubric testing) provided opportunities for ongoing analysis and improvement. Summative evaluation efforts have been designed to provide evidence of usability and effectiveness. This report focuses on the summative evaluation of the Planets and Moons (PM) unit. Given the experimental design (teachers randomly assigned to treatment or control groups) and the abundance of data collected, the majority of the analyses reported are based on quantitative methods; however, a small random sample of teachers were interviewed to provide an in-depth qualitative perspective on the *Seeds/Roots* intervention as well. *Seeds/Roots* uses an integrated approach to teaching science and literacy, and this evaluation provides evidence for the benefit(s) of utilizing an integrated approach in comparison to standard instructional practices in a fourth and fifth grade Planets and Moons unit.

Background on the Treatment

Seeds/Roots is an integrated science-literacy program designed for Grades 2 through 5, partially based on revisions of units in the Great Explorations in Math and Science Program. The *Seeds/Roots* unit is designed as a next generation of standards-aligned elementary inquiry science materials that advance student learning in science while meeting the challenges of an increasingly congested school day, low levels of elementary teacher preparation and efficacy in science, the pressures of large-scale testing, and the growing diversity of our nation's classrooms. *Seeds/Roots* science-literacy integration is based on previous literature on integrated methods. The emphasis is on integrating content-area learning, reading, and writing. This approach to science-literacy integration ideally fosters a synergistic relationship (Cervetti, Pearson, Bravo, & Barber, 2006). The *Seeds/Roots* model builds on previous work that has demonstrated positive effects from using an integrated approach (Guthrie & Ozgungor, 2002; Romance & Vitale, 1992). There are three approaches to instructional integration (Stoddart, Pinal, Latzke, & Canaday, 2002): a thematic approach characterized by the use of overarching themes to create connections among domains, an interdisciplinary approach in which content or processes in one domain are used to support learning in another, and an integrated approach in which emphasis on two or more domains is balanced. Details of *Seeds/Roots'* integrated curriculum and process to achieve balance are discussed in Cervetti, Barber, Dorph, Pearson, and Goldschmidt (2009).

Evaluation Design and Objectives

In order to determine whether there are statistically significant and substantively important effects from using an integrated science and literacy approach to instruction compared to content-comparable business-as-usual science instruction, the *Seeds/Roots* unit

was embedded in a curriculum unit on planets and moons which involved students in talking, reading, and writing about the characteristics of planets and moons. The unit also provided opportunities for explicit instruction of literacy abilities, such as using the reading comprehension strategies of making predictions and summarizing, writing summaries, using nonfiction text structures to find information, and engaging in oral discourse.

During the 2008–2009 school year, approximately 100 fifth grade teachers from about 60 schools, 10 states, and both rural and urban counties participated in the study. The states were selected as study sites because of a consistent relationship between the state's science standards addressing planets and moons and the content of the integrated science-literacy PM unit, more easily enabling a content-comparable comparison group. Teachers were randomly assigned to either (1) present the integrated science-literacy PM unit to their students (treatment group) or (2) present the content of their state science standards related to PM using whatever curriculum materials they regularly use (control group).

LHS researchers administered pretests and posttests in science and literacy to students in all treatment and control classrooms during the weeks before and after a 12-week teaching window. The evaluation plan called for quantitative summative analysis of student performance, student attitudes, teacher attitudes, and teacher efficacy. The plan was intended to evaluate these elements by collecting data using the following instruments for students:

1. An assessment of science knowledge.
2. An assessment of the nature of science.
3. An assessment of science inquiry.
4. An assessment of science vocabulary.
5. An assessment of reading comprehension using related and unrelated science passages.
6. A science writing assessment.
7. An assessment of student attitudes towards science.
8. Student demographics collected from districts as well as their results on the state standardized test results for science and English language arts.

For teachers, the following instruments were used:

1. A survey of teacher background.
2. Pre and postsurveys of teacher attitudes and self-efficacy.

Given these data, the evaluation focused on examining two aspects related to the implementation and effectiveness of the *Seeds/Roots* unit. Evaluation of implementation relates to examining the impact of implementation on outcomes as well as examining teacher

perceptions regarding the unit's efficacy and student engagement. Effectiveness is evaluated by examining outcomes related to student learning in science, student learning in literacy, and teacher attitudes and practices. In addition, student demographic and state achievement information is used to disaggregate and triangulate student results. Disaggregation of results is an important aspect as it presents an opportunity to examine whether the *Seeds/Roots* unit is particularly beneficial for students at risk. In this case, this relates to low socio-economic status free/reduced lunch or Title I students and English Language Learners (ELL). Triangulation of results relates to using independent assessments (i.e., the *Seeds/Roots* unit assessments and state assessments, as well as teacher perceptions of efficacy). Given that students are assigned to treatments by teacher (cluster randomized design), and teachers teach within schools, a multilevel modeling framework is used to account for the design and the lack of independence among observations and to take advantage of the data structure by examining the potential impact of context on treatment effects. The multilevel model (MLM) analyses are outlined below. The following research questions guided the data collection and choice of analyses methods. Broadly, the categories of data collected include (1) student academic outcomes, (2) student affective outcomes, (3) teacher outcomes, and (4) implementation.

Does the *Seeds/Roots* treatment result in higher student performance compared to the business-as-usual condition in science content?

1. **Student Academic Outcomes**
    a. Does the *Seeds/Roots* Planets and Moon curriculum help students make progress in science?
    b. Does the *Seeds/Roots* treatment result in higher student performance in science compared to the business-as-usual condition of using a standard science curriculum?
    c. Does the *Seeds/Roots* curriculum help students make progress in literacy (i.e., vocabulary and reading)?
    d. Does the *Seeds/Roots* treatment result in higher student performance in literacy (i.e., vocabulary and reading) compared to the business-as-usual condition of using a standard literacy curriculum?
    e. Does the *Seeds/Roots* treatment help English Language Learners (ELLs) make similar progress compared to English-Only (EO) students?
    f. Does the *Seeds/Roots* treatment help students in some states more than in other states?

2. **Student Affective Outcomes**
    a. Does the *Seeds/Roots* treatment have an effect on students' engagement and attitude toward science and literacy?
3. **Teacher Outcomes**
    a. Does the *Seeds/Roots* treatment influence teachers' attitude toward science teaching? Toward literacy teaching?
    b. Do teacher background factors affect student outcomes in control and treatment conditions?
4. **Implementation**
    a. What factors distinguish between successful and less successful implementation of the treatment?
    b. What are teachers' reactions to the quality, usability, and utility of the units? What are some positive aspects of the *Seeds/Roots* Planets and Moon unit; what could be improved?

Methods and Data

Methods

In studies of program or intervention effects in schools using pre and posttests, students are typically nested within different sites (classrooms). Ignoring the nested structure of the data gives rise to two main problems: misleadingly small standard errors for treatment effect estimates and failure to detect between-site (classroom) heterogeneity in intervention effects (Raudenbush & Bryk, 2002; Seltzer, 2004; Snijders & Bosker, 1999). The between-site heterogeneity is not surprising; because class intake can vary, teachers can vary considerably in terms of implementation, background characteristics of participants, as well as factors that are related to the treatment effects. This is both a statistically and substantively important issue. By using a three-level random effects model, we are able to divide the variation in achievement into between-student, between-teacher, and error components. This is particularly important to do because data containing multiple levels of aggregation can lead to errors in interpretation when these multiple levels are ignored (Aitkin & Longford, 1986; Burstein, 1980).

We utilize MLMs—specifically, a three-level model that includes students, teachers, and schools. This three-level MLM forms the basis for analyses of the outcomes using various specifications of the following model (Equation 1). The model consists of three levels and allows for a flexible specification of the covariance structure at every level of the analysis (Snijders & Bosker, 1999). MLMs are flexible yet powerful tools for understanding the impact of a treatment on student performance (Raudenbush & Bryk, 2002). In order to

examine the potential impact of the treatment, we use lagged performance in order to examine residual change in student performance. Using a three-level model, students represent Level 1; teachers, Level 2; and schools, Level 3.

The Level 1 model is

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} (Y_{ijk} - Y_{..k}) + e_{ijk} \quad (1a)$$

Y_{ijk} is the outcome (e.g., *Seeds/Roots* science content assessment) for Student *i* in Class[2] *j* in School *k*. π_{0jk} represents the mean outcome of Classroom *j* in School *k* and π_{1jk} represents the relationship between the pre and the posttest. Finally, e_{ijk} is a random student effect.

At Level 2 (between teachers, within schools) we model the impact of the treatment, given that treatment assignment was by teacher (teacher level).

$$\pi_{0jk} = \beta_{00k} + \lambda_{01k} TRT_{jk} + r_{0j}$$

$$\pi_{1jk} = \beta_{10k} + r_{1j} \quad (2)$$

In Equation 2, β_{00k} represents the school mean performance, while λ_{01k} represents the treatment effect. Both r_{0jk} and r_{1jk} are random teacher effects. Using Equation 2 alters the interpretation of π_{0jk}. Now π_{0jk} is the mean class performance of control classrooms and $\pi_{0jk} + \lambda_{01k}$ is the mean performance of treatment classrooms.

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (3)$$

$$\lambda_{01k} = \gamma_{010}$$

$$\beta_{10k} = \gamma_{100}$$

$$\lambda_{11k} = \gamma_{110}$$

In Equation 3, γ_{000} is the grand mean of student performance. γ_{010} is the overall treatment effect.

The Level 1 model represented in Equation 1a can be further specified to account for differences in classroom intake characteristics—for example, pretest performance or student background characteristics. The Level 1 model, then, becomes:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} (Y_{ijk} - Y_{..k}) + e_{ijk} \quad (1b)$$

[2] We use the term class and teacher interchangeably. It is natural to consider a group of students sitting in a classroom, but each classroom is taught by a single teacher. Moreover, student performance is considered to be impacted by the teacher.

Hence, π_{0jk} becomes the adjusted mean outcome of control[3] Classroom *j* in School *k*.

$$\pi_{1jk} = \beta_{10k} + \gamma_{11k} \text{TRT}_{jk} + r_{1jk} \quad (2b)$$

Given the extension (or possible extension) in Equation 1b, the Level 2 model can be reparameterized to include treatment indicators. Hence, as shown in Equation 2b, β_{10k} represents the mean class relationship between the pretest and the posttest in control classrooms. γ_{11k} represents the cross-level interaction between the treatment and pretest scores, whereas γ_{01k} represents the main effect of the treatment—that is, did treatment classrooms outperform control classrooms? Given pretest performance, γ_{11k} estimates whether the treatment is differentially effective for students with different levels of preparedness (i.e., pretest scores). This cross-level interaction tests whether the treatment is differentially more effective for low achievers (when $\gamma_{11k} < 1$) or more effective for high achievers (when $\gamma_{11k} > 1$). This becomes an important mechanism for testing the differential impact of the treatment on specific subgroups of students. The previous example uses prior student knowledge, which allows for the evaluation of the *Seeds/Roots* impact on low/high achievers. Additional student characteristics can be added to Equation 1b and tested by expanding Equation 2b (e.g., including ELL status in Equation 1b and adding a $\gamma_{11k} \text{TRT}_{jk}$ into Equation 2b).

At Level 3, we account for the fact that classrooms are nested within schools. Using an average pretest for the classroom tests the impact of the classroom average achievement, or context, on individual student posttest performance. An interaction between the treatment and control indicator and the average classroom performance tests whether the impact of average classroom performance affects individual student performance differently in control and treatment classrooms.

We generally use equally weighted composites as the metric for analysis, although we do use an item response theory (IRT)-based score using a combination of dichotomous multiple choice and ordered multiple choice, polytomous-scored items. This potentially provides more in-depth information regarding student science content knowledge and subsequent changes in that knowledge, as items can provide indicators of levels of knowledge and responses are based on concept maps (Briggs, Alonzo, Schwab, & Wilson, 2006). The IRT-based scores are moderately correlated to equally weighted composites, the latter being substantively easier to interpret.

[3] Control classroom given in Model 2.

However, using IRT-based scores and their associated standard errors of measurement, we can use a three-level model to model the posttest at *t* and the pretest at *t-1* to explicitly model true student gains. The model is based on the analysis (Bryk, Thum, Easton, & Luppescu, 1998). At Level 1:[4]

$$y_{tijk} = \alpha_{1tijk} + \alpha_{2ijt}\pi_{2ij} + e_{ijt} \quad (4)$$

Here, for Student *i* with Teacher *j* at Time *t*, the assessment scale score is denoted as y_{tij}. Time in this instance refers to the pretest, $t = 0$, and the posttest, $t = 1$. Equation 3 estimates two parameters: student's initial status for the pretest (π_{1ij}) and gain on the posttest (π_{2ij}). Given this parameterization, α_1 is coded as 1, and $\alpha_2 = 0.1$ for the pre and posttest, respectively. The error, e_{tij}, is assumed to be $N \sim (0,\sigma^2)$. This formulation is a basic growth model formulation (which could include a time varying covariate) (Raudenbush & Bryk, 2002). One concern with Equation 4 is the degrees of freedom available to estimate the random effects of interest. In order to model true initial status (pretest) and true gain, Equation 3 is reconceptualized to take advantage of the estimated precision for each score by using the standard error of measure (SEM), s_{ijt}. Hence:

$$y^*_{tij} = y_{tij}/s_{tij}$$
$$\alpha_1{}^*_{tij} = \alpha_{1tij}/s_{tij}$$
$$\alpha_2{}^*_{tij} = \alpha_{2tij}/s_{tij}$$
$$e^*_{tij} = e_{tij}/s_{tij}$$

Equation 1, reparameterized using the SEM, becomes:

$$y^*_{ijt} = \alpha_1{}^*_{ijt}\pi_{1ij} + \alpha_2{}^*_{ijt}\pi_{2ij} + e^*_{ijt} \quad (4b)$$

In this way, e^*_{ijt} $N \sim (0,1)$, π_{1ij}, and π_{2ij} now estimate a student's true initial status and true gain, respectively (Bryk et al., 1998). At Level 2, student covariates are incorporated to account for between-student differences in true initial status and true gains. Hence, at Level 2,

$$\pi_{1ij} = \beta_{10j} + X_{ij}\beta_{11j} + r_{10j}$$
$$\pi_{2ij} = \beta_{20j} + X_{ij}\beta_{21j} + r_{20j} \quad (5)$$

where X_{ij} represents a vector of student covariates. Gender, language, and economic status are examples of student covariates. The between-teacher model is

[4] The growth model is based on a multilevel model (Raudenbush & Bryk, 2002) where test occasions are nested within students, who are nested within teachers. Hence, Level 1 is the within-student model; Level 2 is the between-student model, and Level 3 is the between-teacher model.

$$\beta_{10} = \lambda_{100} + u_{100}$$
$$\beta_{20} = \lambda_{200} + u_{200} \quad (6)$$

The treatment effect is modeled by expanding Equation 5 to include

$$\beta_{10} = \lambda_{100} + \lambda_{101}TRT + u_{100}$$
$$\beta_{20} = \lambda_{200} + \lambda_{201}TRT + u_{200} \quad (7)$$

In the case of Equation 6, λ_{101} and λ_{201} test for pretreatment differences in true student performance on the pretest (which should be 0, given random assignment) and differences in gains between treatment and control students, respectively. Teacher covariates such as background, experience, and implementation can be added to Equation 7.

Limitations. One school (and several teachers) was eliminated from the analysis because in this school, treatment students were purposely placed into treatment classrooms by the principal. The school's performance was significantly below the mean performance of the rest of the sample. Another limitation relates to the analyses of student demographics and state assessment outcomes. Extensive efforts to retrieve state data, post hoc, resulted in a substantial amount of non-response. Preliminary analyses revealed that students for whom we received additional background information were not representative of the sample as a whole, and reporting results based on this subset would obfuscate results and the benefits derived from a cluster randomized trial.[5] Another potential problem is that the study was carried out in ten states—making the control condition against which the treatment is compared quite variable and less structured. This generally makes it more difficult to identify significant treatment effects (Campbell & Stanley, 1963).

Data

The dataset used for analysis consists of data from four sources: (1) individual student assessment pre and postresults based on LHS-generated assessments (these contain individual student observations on several assessment measures, including scored writing assessments); (2) teacher survey results, including responses from both treatment and control teachers; (3) individual student data provided by participating districts;[6] and (4) teacher interview responses. In the following pages, we describe the sample of teachers participating in the study as well as the assessments used to evaluate whether the teachers using the *Seeds/Roots* curriculum demonstrated statistically significant and substantively important effects on science and literacy compared to business as usual.

[5] We conducted preliminary analyses of the subset of data, and results were inconsistent and generally provided no further illumination of treatment or implementation effects.
[6] See Footnote 5.

Teachers. Table 1 presents prestudy background descriptive information related to the participating teachers—this includes teachers who were randomly assigned to the treatment and control groups.[7] Overall, teachers in the treatment and control groups were fairly similar in experience and training, although treatment teachers were more likely to have an Early Childhood Education certification than control teachers were. Conversely, treatment teachers tended to have more education on average, with about 40% having greater than a bachelor's degree—compared to about 30% for control teachers. Before treatment, we evaluated which teachers were inquiry-based teachers based on an internally developed index as well as on the reported time spent on hands-on activities. Both indicators indicate that treatment and control groups were approximately equal in terms of providing inquiry-based instruction, pretreatment. We also examined pretreatment teacher perceptions of related self-efficacy, of the importance of the teacher in learning science, and of science content knowledge. Although responses differ by group, there are no substantively important differences between treatment and control teachers, except that treatment teachers tended to agree that teaching English Language Arts was important at a slightly higher rate than control teachers did (67% vs. 50%).

[7] The results reported in Table 1 exclude teachers that had no student information (dropped before the study collected any data) as well as teachers who were in a school where diffusion/contamination potentially impacts results. One school placed control students into treatment classrooms, which led to dropping one school with five classrooms. Students that were switched scored substantively lower on pretests than fellow control and treatment students (school average was approximately 0.3 standard deviations below all other students).

Table 1

Teacher Prestudy Characteristics, Practices, & Perceptions

Variable	Control			Treatment		
	N	*M*	*SD*	*N*	*M*	*SD*
Teacher education and experience						
Number of certifications	39	1.44	0.79	41	1.46	0.67
Early Child Education certification	42	0.14	0.35	42	0.31	0.47
English Language Learner (ELL) certification	42	0.17	0.38	42	0.10	0.30
Student with Disabilities (SWD) certification	42	0.02	0.15	42	0.05	0.22
Subject-specific certification	42	0.24	0.43	42	0.17	0.38
Other certification	42	0.36	0.48	42	0.36	0.48
Clear credential	42	0.62	0.49	42	0.55	0.50
Life credential	42	0.17	0.38	42	0.33	0.48
Other credential	42	0.10	0.30	42	0.05	0.22
Responsible for English Language Arts (ELA)	42	0.67	0.48	42	0.67	0.48
Degree < BA	18	0.43	0.48	13	0.29	0.44
Degree = BA	12	0.29	0.44	12	0.27	0.48
Degree > BA	12	0.29	0.44	18	0.40	0.49
Years of teaching at current grade level	40	6.86	5.31	41	7.10	6.22
Prestudy teacher practices						
Teacher inquiry instruct scale	39	22.1	3.86	42	21.9	3.81
Teacher inquiry based	42	0.57	0.50	42	0.55	0.50
Teacher perceptions preunit						
Pretest Science Efficacy (summed)	40	58.58	7.64	42	62.12	8.21
Pretest Literacy Efficacy (summed)	37	56.19	7.25	41	57.51	6.44
Teaching ELA important	42	0.50	0.51	42	0.67	0.48
Teacher effectiveness influence student sci performance	42	0.36	0.48	42	0.31	0.47
Science Content Knowledge	40	10.65	1.61	42	10.67	2.25

Table 2 presents results related to teacher perceptions about elements related to the unit. Sampled classroom characteristics were quite similar. Teacher practices during the unit varied substantially on some aspects. Most notable is that treatment teachers spent approximately 50% more instructional time on the unit than control teachers per week.[8]

[8] An increase in time for the treatment group would be expected given the integrated nature of the treatment.

Among reported practices, however, treatment and control teachers tended to allocate their time in approximately the same manner—implying that treatment students received more instruction across the listed elements. In both treatment and control classrooms, about one third of the time was spent on hands-on activities. Substantively, this implies that control students spent about an hour on hands-on activities during the week, while treatment students spent about an hour and 20 minutes conducting hands-on activities. The difference in writing was more substantial, with control students spending about 25 minutes on writing and treatment students spending about 42 minutes on writing. In both groups, about 15% was spent on writing, but slightly more time was spent reading textbooks in treatment classrooms (20% vs. 17%).

Table 2

Teacher During-Study Characteristics and Practices

	Control			Treatment		
Variable	*N*	*M*	*SD*	*N*	*M*	*SD*
Classroom characteristics						
Total number of students in classroom	42	22.97	4.55	42	23.19	4.16
How many students are ELL	42	4.93	7.71	42	3.14	6.24
Percent of student in classroom are ELL	42	21.11	32.06	42	14.57	29.71
During study teacher practices						
Science instruction minutes per week	35	180.2	76.04	38	270.0	97.41
Hands-on inquiry (% of sci instructional time)	35	33.71	18.60	38	30.66	15.47
Reading books/textbooks (% of sci instructional time)	35	17.29	9.02	38	20.39	8.65
Class discussions (% of science instructional time)	35	22.29	11.84	38	22.50	9.98
Writing (% of science instructional time)	35	14.11	7.32	38	15.61	6.85
Science vocabulary (%)	35	12.29	5.33	38	11.50	5.05
Number of lessons taught	42	29.41	11.70	42	31.23	8.84

Table 3 presents postunit perceptions as well as postunit perceptions specifically related to the *Seeds/Roots* unit—only answered by treatment teachers. Teacher responses related to self-efficacy were again substantively similar between treatment and control teachers. Results related to the PM unit showed that treatment and control teachers had relatively similar perceptions regarding the PM unit, with one exception. About two thirds of teachers using the *Seeds/Roots* curriculum indicated that they spent more time than usual on the PM unit, while only about one fifth of control teachers indicated that they spent more time than

usual on PM. Consistent with expectations, the control teachers thought that the unit supported standards at a slightly greater rate than treatment teachers (71% vs. 62%), but the fact that the *Seeds/Roots* curriculum was implanted in 10 states indicates that the unit is fairly transferrable. It is interesting to note that teachers within states did not all uniformly agree whether the unit supported standards or not. It is also interesting to note that, in both groups, roughly half of the teachers thought the challenge level was inappropriate. Also, teachers in both groups felt the unit was more effective for high achievers than for low achievers or ELL students.

Table 3

Teacher Poststudy Perceptions

	Control			Treatment		
Variable	*N*	*M*	*SD*	*N*	*M*	*SD*
Teacher perceptions post unit						
Posttest science efficacy (summed)	36	60.78	7.18	38	62.97	8.22
Posttest literacy efficacy (summed)	32	57.47	6.96	38	56.82	6.60
Unit is engaging	42	0.64	0.48	42	0.71	0.46
Unit challenge level appropriate	42	0.55	0.50	42	0.48	0.51
Unit supports standards well	42	0.71	0.46	42	0.62	0.49
Implementation successful	42	0.60	0.50	42	0.48	0.51
Unit effective for English Language Learners	42	0.38	0.49	42	0.36	0.48
Unit effective for low achievers	42	0.31	0.47	42	0.40	0.50
Unit effective for high achievers	42	0.76	0.43	42	0.74	0.45
Spent more time on unit than normal	42	0.21	0.42	42	0.69	0.47
Teacher perceptions related to treatment						
Which materials are more engaging for students?[a]				29	0.76	0.44
Which materials better support teaching?[a]				31	0.77	0.43
Teacher comfortable with *Seeds/Roots*				35	0.51	0.51
Seeds/Roots material different				33	0.64	0.49
Use *Seeds/Roots* material again				35	0.60	0.50

Note. [a] *Seeds/Roots* vs. regular.

Table 3 also summarizes treatment teachers' perceptions related specifically to the *Seeds/Roots* PM unit. While about three-fourths of the treatment teachers felt the *Seeds/Roots* unit was more engaging and supported teaching better than the standard unit, only about half

of the teachers felt comfortable with the unit. Still, about 60% indicated that they would likely use the unit again.[9]

Assessments. Table 4 presents the reliabilities of the pre and postassessments developed by LHS. An assessment's reliability represents score consistency for individual students. However, the reliability of classroom or teacher assessment means provides an indication of how well we can distinguish among classrooms in true student performance. A low reliability for an assessment is generally substantially higher when aggregated to the classroom level. However, low assessment reliability significantly impacts the reliability of gain scores. Hence, gain scores potentially obfuscate the impact of the treatment. Generally, the reliabilities displayed in Table 4 are acceptable (or close to the generally accepted criterion of 0.70).

Table 4

Reliability of Science Assessment

Assessment	*n* items	*Cronbach's* α
Pretreatment		
Reading	10	0.67
Vocabulary	25	0.75
Science Content	30	0.78
Science Inquiry	14	0.69
Nature of Science	12	0.66
Posttreatment		
Reading	10	0.66
Vocabulary	25	0.81
Science Content	30	0.82
Science Inquiry	14	0.73
Nature of Science	12	0.69

Table 5 summarizes the descriptive results for the five assessments given to students prior to the unit and after the unit. Using equally weighted composite scores as the primary metric for evaluation (except for science content discussed in the results section, which also uses IRT-based ability estimates as outcomes), the total possible for each assessment corresponds to the number of items presented in Table 4. Hence, the average pretest score in

[9] We present more detail related to this when addressing Question 4.

science content represents students scoring about 39% correct on the pretest and about 51% correct on the posttest. Students performed substantively better on the other assessments—generally answering about two thirds to three fourths of the items correctly.

Table 5

Descriptive Results for Science Assessment

	Total			Control			Treatment		
Label	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Vocabulary Pretest	1,886	15.30	4.26	913	15.18	4.12	973	15.41	4.38
Vocabulary Posttest	1,593	18.18	4.46	790	17.34	4.28	803	19.01	4.49
Reading Pretest	1,885	6.40	2.21	913	6.38	2.22	972	6.41	2.20
Reading Posttest	1,591	6.95	2.17	789	6.90	2.17	802	7.00	2.18
Science Content Pretest	1,889	11.59	5.18	914	11.62	5.18	975	11.57	5.19
Science Content Posttest	1,590	15.53	5.86	788	14.82	5.91	802	16.22	5.74
Science Inquiry Pretest	1,870	9.17	2.87	904	9.28	2.79	966	9.06	2.93
Science Inquiry Posttest	1,587	10.27	2.80	775	10.08	2.86	812	10.45	2.74
Nature of Science Pretest	1,549	8.25	2.53	731	8.31	2.58	818	8.21	2.49
Nature of Science Posttest	1,370	9.19	2.50	656	8.95	2.69	714	9.42	2.30

Table 6 provides summary statistics for the five writing domains as well as the estimated reliability for an overall writing score. The reliabilities for the overall writing scores are both acceptable. The Concepts and Evidence domains are scored on a four-point scale, while the remaining three domains are scored on a three-point scale. A cursory examination of Table 6 indicates that there was no improvement in students' introduction and conclusion writing; however, there appear to be small gains in both science concepts and evidence. Table 7 presents the number of lessons that were completed by the treatment and control teachers. It is important to note that the number of lessons is not directly comparable between treatment and control teachers, especially given that the study was carried out in 10 states and that lesson content differed by definition across condition. The standard deviations are quite high, indicating substantial variation in the number of lessons completed.

Table 6

Descriptive Results for Five Writing Domains

	Total			Control			Treatment		
Variable	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Science concepts – pre	401	1.78	0.62	193	1.77	0.61	208	1.78	0.64
Science concepts – post	398	2.05	0.72	189	1.95	0.69	209	2.13	0.73
Evidence – pre	401	1.78	0.65	193	1.79	0.63	208	1.76	0.67
Evidence – post	398	1.90	0.64	189	1.90	0.65	209	1.90	0.63
Cohesion – pre	401	1.76	0.57	193	1.78	0.54	208	1.74	0.60
Cohesion – post	398	1.83	0.51	189	1.83	0.51	209	1.84	0.52
Introduction – pre	401	1.95	0.75	193	1.98	0.74	208	1.92	0.77
Introduction – post	398	2.00	0.72	189	1.95	0.72	209	2.05	0.71
Conclusion – pre	401	1.23	0.58	193	1.23	0.59	208	1.23	0.58
Conclusion – post	398	1.28	0.62	189	1.24	0.60	209	1.31	0.64
Reliability (Cronbach's α)									
Pretest	0.77								
Posttest	0.74								

Table 7

Number of Science Lessons

Completed	Control	Treatment
M	29.41	31.22
SD	11.70	8.74
N	42	43

Evaluation Results

In evaluating the *Seeds/Roots* curriculum on student outcomes, we address each of the research questions in turn. The analyses often consist of multiple tables and, in some instances, results may not unequivocally answer the evaluation questions.

Student Academic Outcomes

1a) Does the Seeds/Roots Planets and Moon curriculum help students make progress in science? Table 8 summarizes treatment student progress on the three science

assessments. The results indicate that students demonstrated significant improvement in all three domains. On all three assessments, student improvement was approximately one standard deviation from pre to posttest.

Table 8

Science Pre-Post Gains

	N	*M*	*se*	*signif*
Science Content[a]	759	4.45	0.18	***
Inquiry Science	731	1.40	0.09	***
Nature of Science	627	1.17	0.09	***

Note. [a]Science content gain based on mean difference of dichotomously scored items.
* $p < .10$. ** $p < .05$. *** $p < .01$

1b) Does the *Seeds/Roots* treatment result in higher student performance in science compared to the business-as-usual condition of using a standard science curriculum? We first estimate treatment effects on the three science-related outcomes (content, nature of science, and science inquiry) but focus our attention on science content, as this is the primary indicator related to the unit's content and also has better reliability than the other, significantly shorter, assessments. The results in Tables 9 through 11 summarize the treatment control comparisons using the MLM models described. We note again that the quantitative results are based on 86 teachers (42 control and 44 treatment) who had student data and who were not excluded due to contamination. The results in Tables 9a and 9b pertain to science content and indicate that, when considering posttest scores and controlling for pretest performance, treatment students scored about 1.3 points higher than control students. This difference fails to provide sufficient evidence that we can reject the null hypothesis at the traditional .05 level but does provide suggestive evidence that the treatment has a systematic impact on student performance ($p < .10$). The results in Model 2 indicate that the treatment is no more or less effective for low- or high-achieving students (based on the pretest). That is, there is no joint or interaction effect between the treatment condition and pretest performance.

Table 9a

Estimated Treatment Effects on Student Science Content Posttest Results

Model	Science Content			
	Model 1	*signif*	Model 2	*signif*
Fixed Effects				
Mean Posttest				
Control classroom	14.85		14.94	
Treatment classroom	16.13	^	16.13	^
Treatment effect size				
Pretest	0.60	***	0.59	***
Treatment x Pretest (cross-level) interaction			0.03	
Treatment effect size				
Random Effects (Variance component)				
Level 1: Student	25.42		18.39	
Level 2: Intercept	9.01	***	4.54	***
Level 2: Slope (pre science content)				

Note. $N = 1{,}590$.
$^p < .10$. $*p < .05$. $** p < .01$. $***p < .001$.

Another way to examine whether there was a treatment effect is to compare gains made by comparison classrooms and treatment classrooms. It is important to note that this addresses a slightly different research question than the one originally posed in 1b. Also, using gains does not account for pretreatment differences in comparison and treatment classrooms and may be important to consider along with bias associated with attrition.[10] The results using gain scores are presented in Table 9b. The results indicate that there is significant between-teacher variability,[11] which necessitates the use of some mechanism to estimate correct standard errors, here accomplished by again using an MLM. The results presented in Table 9b indicate that students in treatment classrooms gained about 1 point more than students in comparison classrooms.

[10] It was not possible to calculate differential attrition rates for treatment and control classrooms as data on teachers who dropped at initial project stages was not reliable in indentifying condition.
[11] The unconditional model for gains indicates that ICC is .17 ($p < .01$)

Table 9b

Estimated Treatment Effects on Student Science Content Pre-Post Gains Model	Science Content		
	Model 1	*se*	*signif*
Fixed Effects			
Mean Post-test			
Control Classroom Gains	3.35	0.42	***
Treatment on Gains	1.08	0.52	*
Random Effects (Variance component)			
Level 1: Student	2.02		***
Level 2: teacher means	4.66		

Note. df 76 teachers
^$p < .10$. *$p < .05$. ** $p < .01$. ***$p < .001$.

Table 10

Estimated Treatment Effects on Student Nature of Science Posttest Results

Model	Nature of Science			
	Model 1	*signif*	Model 2	*signif*
Fixed Effects				
Mean Posttest				
Control classroom	9.01	***	8.97	
Treatment classroom	9.41	*	9.45	*
Treatment effect size	0.16		0.19	
Pretest	0.52	***	0.57	***
Treatment x Pretest (cross-level) interaction			-0.10	^
Treatment effect size				
Random Effects (Variance component)				
Level 1: Student	1.96		1.96	
Level 2: Intercept	0.67	***	0.66	***
Level 2: Slope (pre Nature of Science)	0.13	***	0.12	***

Note. N = 1,593.
^$p < .10$. *$p < .05$. **$p < .01$. *** $p < .001$.

The results in Table 10 indicate that students in treatment classrooms score about 0.40 points higher on the posttest, accounting for pretest performance on the nature of science

assessment ($p < .05$). The results in Model 2 indicate that there is suggestive evidence that the *Seeds/Roots* curriculum is more effective for low-achieving students (based on the pre nature of science assessment) than for high-achieving students. However, given the small (main) treatment effect, the estimated interaction (though not statistically significant) implies that treatment would benefit only the lowest (in terms of pretest performance) 10% of students on the nature of science assessment.

Table 11 presents results for the science inquiry results. Consistent with the science content results, there is no treatment effect (although, if the cross-level interaction is included, some suggestive evidence emerges). That is, accounting for the pretest performance and the potential differential impact of the treatment on students at different levels of pretest intake, students in the treatment condition score about 0.4 points higher on the posttest ($p < .10$).

Table 11

Estimated Treatment Effects on Student Science Inquiry Posttest Results

	Science Inquiry			
Model	Model 1	*signif*	Model 2	*signif*
Fixed Effects				
Mean Posttest				
Control classroom	10.18	***	10.11	***
Treatment classroom	10.44		10.50	^
Treatment effect size				
Pretest	0.56	***	0.60	***
Treatment x Pretest (cross-level) interaction			-0.07	
Treatment effect size				
Random Effects (Variance component)				
Level 1: Student	2.11		2.10	
Level 2: Intercept	.10	***	.09	***
Level 2: Slope (pre Inquiry)	.75	*	.75	^

Note. $N = 1{,}587$.
^$p < .10$. *$p < .05$. ** $p < .01$. ***$p < .001$.

We next investigate whether using more advanced scoring options and modeling changes in student science content ability provides additional insight into potential treatment effects. As noted in the modeling section, using IRT scores has several benefits and allows us

to eliminate potential spurious relationships between initial status and growth (that observed gains exhibit).[12] LHS provided ability estimates based on the same science content assessment examined in Table 9a, but 10 of the polytimous items were rescored to potentially provide more in-depth information regarding student performance. As noted, the IRT ability estimates are moderately correlated with raw, equally weighted scores. The results of the MLM models described in Equations 8 through 10 provide consistent results in that the null hypothesis related to a treatment effect cannot be rejected. IRT model results are consistent with multilevel gain model results. The results presented in Table 12 imply that students did not demonstrate gains from pre to posttests; there was a slight decrease in performance (about 0.06 standard deviations). The suggestive results indicate that students in treatment classrooms demonstrated a smaller drop in performance than control students. However, the model fit is very poor and is based on the change in the deviance.

Table 12

Estimated Treatment Effect Using IRT Scores[a]

	Estimate	*signif*
Fixed Effects		
Science Pretest		
Control classroom	0.014	
Treatment classroom	0.011	
Science Posttest gain		
Control classroom	-0.044	**
Treatment classroom	-0.017	^
Treatment Effect Size		
Random Effects	SD	signif χ^2
Between student		
Pre-post gain	0.006	>.500
Between teacher		
Science pretest	0.093	< .001

Note. N = 1,609.
[a]Includes polytimous scored items.
^p < .10. *p < .05. **p < .01. ***p < .001.

[12] We include these analyses because, previously, these models corroborated simple raw score MLM models and demonstrated that results were robust to different modeling options. However, in this case, due to the exploratory nature of the polytimous IRT models, we consider these results less reliable but present them for completeness.

1c) Does the *Seeds/Roots* curriculum help students make progress in literacy (i.e., vocabulary and reading)? We next address additional student academic outcomes. Table 13 demonstrates that students receiving the *Seeds/Roots* curriculum exhibited significant growth in reading and vocabulary as did students in control classrooms ($p < .05$).

Table 13

Reading and Vocabulary Gains

Condition	Reading	Vocabulary
Control		
M	0.60	2.40
SD	2.02	3.20
N	725	726
Treatment		
M	0.64	3.40
SD	2.12	3.91
N	758	759

1d) Does the *Seeds/Roots* treatment result in higher student performance in literacy (i.e., vocabulary and reading) compared to the business-as-usual condition of using a standard literacy curriculum? Table 14 presents results for vocabulary and reading. The results indicate that students in treatment classrooms scored about 1.6 points higher than students in control classrooms, accounting for pretest performance. This is an effect size of approximately 0.38. There were no statistically significant differences between treatment and control students in reading.

Table 14

Estimated Treatment Effects on Student Posttest Results

	Vocabulary				Reading			
	3		4		5		6	
Model	Coef.	*signif*	Coef.	*signif*	Coef.	*signif*	Coef.	*signif*
Fixed Effects								
Mean Posttest								
Control classroom	17.34		17.70		6.89		6.98	
Treatment classroom	18.94	**	19.03	***	6.95		7.00	
Treatment effect size	0.38							
Treatment x Pretest (cross-level) interaction			-0.10				-0.01	
Treatment effect size								
Random Effects (Variance component)								
Level 1: Student	14.28		8.97		3.96		2.92	
Level 2: Intercept	5.29	***	1.65	***	0.85	***	0.19	***
Level 2: Slope			0.02	***			0.01	**

Note. $N = 1{,}483$.
*$p < .05$ ** $p < .01$. ***$p < .001$.

We next examine the impact of the *Seeds/Roots* curriculum on student writing. A subset of student essays was scored on five dimensions (listed in Table 6). Table 15 presents the correlations among the scored domains. Overall, only science concepts and evidence scores are moderately correlated, while the remaining domains have low to moderate correlations. This implies that each of the scored domains taps into a different aspect of the student writing.

Table 15

Correlations Among Writing Dimensions

	Pretest (n = 401)			
	Evidence	Cohesion	Introduction	Conclusion
Science Concepts	0.657	0.515	0.332	0.233
Evidence	1.000	0.548	0.369	0.289
Cohesion		1.000	0.479	0.389
Introduction			1.000	0.268
Conclusion				1.000
	Posttest (N = 398)			
Science Concepts	0.571	0.436	0.274	0.242
Evidence	1.000	0.527	0.399	0.228
Cohesion		1.000	0.522	0.294
Introduction			1.000	0.281
Conclusion				1.000

Consistent with previous analyses of *Seeds/Roots* effects on student writing (Goldschmidt & Jung, 2009), there are two possible avenues to proceed: (1) to examine the underlying latent writing achievement based on the observed scores on the five dimensions, and (2) to examine student achievement based on each domain separately. Ultimately, in order to determine whether the treatment had a significant effect on student writing, the former is more appropriate as it controls for the intraperson correlation of scores; however, the latter provides more information in that different results for separate domains can provide additional formative information.

In order to test the global research hypothesis as to whether the *Seeds/Roots* unit results in statistically significant and substantively higher outcomes than the control, the former model is tested. The results are presented in Table 16. The results in Table 16 indicate that there is no difference in writing performance between the treatment and control groups on the pretest, that overall there was no gain in performance ($p > .05$), and that treatment and control students continued to demonstrate similar writing performance on the postunit administration.

Table 16

Estimated Treatment on Latent Student Writing Results

	Writing estimate	*signif*
Fixed Effects		
Mean pretest		
Control classroom	1.76	***
Treatment classroom	1.74	
Mean posttest		
Control classroom	1.84	^
Treatment classroom	1.88	
Random Effects		
Heterogeneous random effects		

Note. $N = 398$.
^$p < .10$. *$p < .05$. **$p < .01$.***$p < .001$.

We next examine the changes in each of the writing domains separately. Table 17 presents results for each of the five writing domains. The results for science concepts indicates that treatment students' post writing science concepts score is about 0.2 standard deviations higher than control students' post science concepts writing score.

Table 17

Estimated Treatment on Student Writing by Dimension

Dimension	Fixed Effect	Coefficient	*se*	approx *p*	Effect Size
Science Concepts					
	Control classroom	1.93	0.06	0.00	
	Treatment effect	0.14	0.08	0.00	
	Pretest	0.27	0.05	0.095*	0.20
Evidence					
	Control classroom	1.90	0.05	0.00	
	Treatment effect	-0.03	0.07	0.00	
	Pretest	0.28	0.05	0.63	
Cohesion					
	Control classroom	1.82	0.04	0.00	
	Treatment effect	-0.02	0.06	0.00	
	Pretest	0.25	0.04	0.75	
Introduction					
	Control classroom	1.94	0.06	0.00	
	Treatment effect	0.04	0.09	0.00	
	Pretest	0.25	0.04	0.65	
Conclusion					
	Control classroom	1.26	0.05	0.00	
	Treatment effect	0.00	0.06	0.00	
	Pretest	0.23	0.05	0.97	

Note. $N = 398$.
*p < .10. **p < .05. ***p < .01.

In all writing domains, and consistent with expectations, predomain scores are statistically significantly related to postwriting scores. However, it is only in science concepts that treatment students demonstrate improved writing over control students. Specific science vocabulary use was counted in the pre and postwriting samples. Table 18 summarizes the vocabulary count by condition. The maximum total number of vocabulary words possible on both the pre and the posttest was 29. Students in both conditions tended not to use the ascribed vocabulary, nor was there much improvement from pre to postassessment.

Table 18

Vocabulary Use by Condition

	Writing Vocabulary Count	
Condition	Pre	Post
Control		
M	2.17	2.64
N	192	190
SD	1.24	1.41
Treatment		
M	2.26	2.67
N	202	202
SD	1.14	1.29

To further examine the concept of integrating science and literacy, we evaluate whether overall preparedness in the three assessed domains (science, vocabulary, and reading[13]) relate to postassessment results and whether there is any transfer between reading and science. Hence, the following analyses examine the effect of including all pretest scores and all gain scores. The pretest scores capture a broader picture of student intake, while gains capture the extent to which students can transfer skills and knowledge from one domain to another. Table 19 presents results examining science content. Model 1 indicates that post science content outcomes are related to not only pre science content knowledge, but also pre vocabulary and reading performance. Model 2 indicates that after taking preperformance into account, as well as the potential cross-level interaction of treatment by pretest score, treatment students perform significantly better than control students on post science content. This is despite the fact that there is no statistically significant treatment by pretest effects—including this interaction in the model, which appears to more finely partition the variation in main treatment effects, allowing us to detect a significant treatment effect. In other words, if we account for students' pre science content, vocabulary, and reading levels, as well as the potential interplay of these levels, and being in a treatment classroom we observe a treatment effect. For exploratory purposes, we further expand the specification of Model 1. The information in Model 2 allows us to test whether prevocabulary knowledge might have a greater impact in treatment than in control classrooms, hence providing some guidance as to what type of intake knowledge might influence the impact of the treatment. Although

[13] Writing results are excluded as only a subset of students has all four sets of scores.

accounting for these differences does indicate that there is a main treatment effect, no systematic results are related to either reading or vocabulary in treatment classrooms. That is, there is no extra benefit of being in a treatment and having better vocabulary at the beginning of the unit.

Table 19

Science Posttest Outcome

Fixed effect	Model 1	*signif*	Model 2	*signif*
Main treatment effect	0.84		1.16	*
Science pretest effect	0.43	***	0.43	***
Treatment effect			-0.02	
Vocabulary pretest effect	0.30	***	0.22	**
Treatment effect			0.15	
Reading pretest effect	0.43	***	0.43	***
Treatment effect			-0.01	

Note. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 20 repeats the analyses presented in Table 19, but it uses post reading as the outcome. The results indicate that post reading is related to preperformance levels of all three tested domains but that treatment and control students do not perform differently.

Table 20

Reading Posttest Outcome

Fixed effect	Model 1	*signif*	Model 2	*signif*
Main treatment effect	-0.04		-0.01	
Science pretest effect	0.03	**	0.04	**
Treatment effect			-0.02	
Vocabulary pretest effect	0.12	***	0.12	***
Treatment effect			-0.01	
Reading pretest effect	0.38	***	0.37	***
Treatment effect			0.01	

Note. *$p < .05$. **$p < .01$. ***$p < .001$.

1e) Does the *Seeds/Roots* treatment help English Language Learners (ELLs) make similar progress compared to English-Only (EO) students? In terms identifying treatment effects, using pretests to account for potential differences between treatment and control students not addressed through randomization is adequate to ensure that treatment effects are consistent. However, including student background characteristics enables us to address whether the treatment is potentially effective in closing achievement gaps for various student subgroups. As noted previously, individual student data provided by the districts were not amenable to separate analysis;[14] however, we have available two indicators of English Language Learner (ELL) status: an individual indicator for about 100 students, collected during the initial preassessments, and another aggregate source based on teacher responses regarding their classroom composition. We utilize both to examine the impact of the *Seeds/Roots* curriculum on this subset of students. We can examine whether ELL student performance is systematically different from non-ELL student performance in treatment and control classrooms. We first use the same MLM models as above to examine the impact of *Seeds/Roots* on ELL student performance, based on individual identification of ELL status. We then use teacher-provided classroom composition information to examine the impact of the treatment on ELL students. Using class composition potentially suffers from ecological fallacy, but the analysis below takes advantage of the fact that nine classrooms consisted of

[14] See Footnote 5.

100% ELL students, which provides some leeway in interpreting results. We reanalyze the data, conducting the analyses for mixed classrooms and ELL-only classrooms.

Table 21 presents the results of the models[15] using science, reading, vocabulary, and writing[16] as outcomes and using individual student ELL classification. Overall, the results are consistent with previous results presented above. In each outcome except writing, pretest results are significantly related to posttest results. Table 21 also indicates that, except in writing, non-ELL students in treatment classrooms performed as well as non-ELL students in control classrooms. For example, in science content, English Only (EO) students in control classrooms are expected to score about 15.2 points, while EO students in treatment classrooms are expected to score about 16.5 points. Writing results indicate that EO students in treatment classrooms did not perform as well as EO students in control classrooms, ceteris paribus. The results in Table 21 also indicate that there was a significant performance gap between ELL and EO students in science content and writing in control classrooms and that this gap was larger in treatment classrooms in science (though not statistically significantly so). The ELL-EO performance gap in writing was smaller in treatment classrooms (though not statistically significantly so). While there was no posttest vocabulary ELL-EO performance gap in control classrooms, there is suggestive evidence ($p < .10$) that the difference between treatment and control classrooms was different for ELL and EO students. This difference is about three points (about one standard deviation). Overall, the treatment effect for ELL students is about 0.01 points in science content, 0.62 in reading, -1.4 in vocabulary, and -1.6 in writing. None of these differences is statistically significant, however.

[15] Conducting the analysis on a small subset of students clearly reduces power, but the analyses are informative nonetheless.

[16] Writing analysis sample size was limited to a subset of the 95 students for whom we had individual ELL information. This sub-sample consisted of 23 students.

Table 21

Effect of ELL Status on Performance

	Science Content		Reading		Vocabulary		Writing	
	Estimate	*signif*	Estimate	*signif*	Estimate	*signif*	Estimate	*signif*
Control Classroom (non ELL)	15.16	***	6.89	***	17.29	***	10.63	***
Treatment classroom (non ELL)	16.52		6.55		18.95		8.15	*
Pretest	0.64	***	0.53	***	0.63	***	0.12	
ELL vs. non-ELL in control	-2.50	*	-0.31		0.13		-1.43	*
Effect of treatment on ELL	-1.35		-0.28		-3.07	^	0.85	

Note. $N = 98$.
^p < .10. *p < .05. **p < .01. ***p < .001.

We take advantage of additional information related to ELL classification by using teacher response to the question that asked what percent of the students in the classroom was ELL. This allows classifying students in two ways: (1) individual student data that classified students as ELL or not ELL and (2) teacher reports that 100% of students are ELL—from which we assume that individual students are ELL. Using these methods, we have a total of 227 ELL students on which to conduct the analysis. Of these ELL students, 132 are control and 95 are treatment.

Table 22

Effect of Treatment on ELL Students in High-Percent[a] ELL Classrooms

Outcome	Estimate[b,c]	*se*	*p value*
Science Content	2.76	1.83	0.16
Nature of Science	0.09	0.66	0.89
Science Inquiry	0.57	0.91	0.55
Reading	0.22	0.48	0.66
Vocabulary	0.85	1.16	0.49
Writing	0.21	0.62	0.75

Note. $N = 227$.
[a]A high-percent ELL classroom has ≥20% ELL students. [b]Model accounts for corresponding pretest. [c]Exploratory analyses revealed that teacher perceptions with respect to whether the unit is effective for ELLs demonstrated significant effects (p < .10) in nature of science, reading, and vocabulary (1.4, .8, and 1.9 point higher outcome, respectively).

Results presented in Table 22 indicate that, among ELL students, there was no statistically significant difference between treatment and control ELLs in high-percent ELL classrooms. In science content, for example, the difference among treatment and control ELL students in high-percent ELL classrooms is about 2.76 points, but this difference is not statistically significant.

We next examine the impact of *Seeds/Roots* in ELL-only classrooms.[17] Table 23 summarizes the distribution of ELL-only classrooms between treatment and control conditions. ELL-only classrooms were approximately equally split between treatment and control groups. We present the ELL classroom results in Tables 29 through 32 as the results are based on class aggregate information and amenable to presentation in the context of unit implementation (which is, of course, also a classroom level variable).

Table 23

Distribution of ELL Classrooms

	Less than 100% ELL	ELL only (100% ELL)	Total
Control	37	5	42
Treatment	40	4	44
Total	77	9	86

1f) Does the *Seeds/Roots* treatment help students in some states more than in other states? We next conduct additional exploratory analyses that take advantage of the fact that the study was conducted in 10 states. Table 24 presents the distribution of teachers by state and treatment condition. The majority of teachers came from four states (AZ, TX, LA, and SC). Exploratory multilevel models generated state-level treatment effect estimates. The state-level estimates are presented in Table 25.

[17] None of the students with individual ELL codes attended the ELL-only classrooms.

Table 24

State by Treatment/Control

	Frequency		
State	Control	Treatment	Total
AZ	12	11	23
CO	3	1	4
CT	5	4	9
FL	2	4	6
GA	1	3	4
LA	5	6	11
MO	2	4	6
NJ	0	1	1
SC	5	5	10
TX	7	5	12
Total	42	44	86

It is interesting to note that teachers varied considerably within states on survey responses, including responses that would seem to be constant. For example, teachers in some state were split on whether or not the unit (either the standard or the *Seeds/Roots* unit) addressed state standards well. Treatment effects do not vary significantly, but state mean pretest performance does, even with the restricted number of degrees of freedom. The results in Table 25 provide some exploratory evidence that context potentially plays an important role in impacting the success of an intervention not specifically designed for each site. For example, in science content, CO, LA, and TX had large positive average treatment effects. GA and LA had large average treatment effects in vocabulary. Reading treatment effects were also large in GA and LA. Overall, the state results are speculative but do provide some insight into the role of context.

Table 25

State-Level Treatment Effect Estimates

	Treatment Effects		
State	Science	Vocabulary	Reading
AZ	-0.43	1.32	-0.40
CO	2.68	0.87	-0.37
CT	0.42	1.69	0.02
FL	0.37	0.40	-0.41
GA	0.14	2.57	0.96
LA	2.48	2.76	0.65
MO	0.44	1.71	-0.47
NJ	—	—	—
SC	0.46	1.61	0.07
TX	2.20	0.04	-0.12

Student Affective Outcomes

2) Does the *Seeds/Roots* treatment have an effect on students' engagement and attitude toward science and literacy? Among control teachers, teachers report that 64.3% of students were engaged and 17% were very engaged. Among treatment teachers, the survey indicates that 69.6% and 23% were engaged and very engaged, respectively. Neither the engaged nor the very engaged percentages were statistically different between treatment and control classrooms. Exploratory MLM models revealed that student engagement (as perceived by the teachers) was not predictive of student performance on posttests.

Students' attitude towards science was also assessed pre and postunit for treatment and control students. Overall, students demonstrated a 1.2 point (0.13 standard deviation) gain in attitude ($p < .05$). Table 26 summarizes the results regarding changes in student attitudes towards science and whether treatment students demonstrated a greater change in attitude than control students. The results indicate that post science attitudes were similar between treatment and control classrooms when accounting for initial attitudes towards science. While there was a slight increase in attitudes towards science, overall, the *Seeds/Roots* curriculum did not enhance this change.

Table 26

Estimated Treatment Effects on Student Attitudes Towards Science

	Model			
	1		2	
Fixed Effects				
Mean posttest				
Control classroom	25.71	***	25.60	***
Treatment classroom	25.84		25.96	
Treatment effect size				
Preunit attitudes	0.66	***	0.62	***
Treatment x Pretest (cross-level) interaction			0.08	
Treatment effect size				
Random Effects (Variance component)				
Level 1: Student	6.51		6.51	
Level 2: Intercept	2.26		2.26	
Level 2: slope (pre-Att)	0.15		0.15	

Note. N = 1,248.
^$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Teacher Outcomes

3a) Does the *Seeds/Roots* treatment influence teachers' attitudes toward science teaching? Toward literacy teaching? Teachers in both conditions were given a self-efficacy survey designed to assess each teacher's perceived self-efficacy in teaching science and literacy. The survey was administered prior to the PM unit and after the PM unit. Previously, we reported that preunit self-efficacy was substantively similar between treatment and control classrooms. Table 27 presents how teacher self-efficacy changed over the course of the unit. Changes in self-efficacy were relatively minor, and there were no significant difference in changes between the treatment and control teachers.

Table 27

Change in Teacher Self-Efficacy

Content	Group	*N*	Mean Change	*SD*	*se*	Difference	Diff.*se*
Science	Control	34	1.29	4.55	0.78		
	Treatment	38	1.00	4.72	0.77	-0.29	1.10
Literacy	Control	31	1.00	4.34	0.78		
	Treatment	37	-0.32	4.26	0.70	-1.32	1.05

3b) Do teacher background factors affect student outcomes in control and treatment conditions? Based on teacher survey responses, we examined the impact of teacher characteristics on student outcomes and in relation to the treatment. The variables we examined included the following:

- Background
  - Credential type
  - Number of credentials
  - Certification level
  - Years of teaching experience
  - Number of certifications
  - Number of times taught Planets and Moons
  - Degree earned
  - Self-efficacy (appropriate for outcome—science or literacy)
- Teacher practices
  - Percent of time spent on hands-on experiences
  - Percent of time spent on reading
  - Percent of time spent on writing
  - Percent of time spent on class discussions
  - Percent of time spent on vocabulary
  - Hours of science instruction
  - Hours of literacy instruction
  - Minutes taught science previously
  - Minutes of science instruction this unit
  - Responsible for science and literacy
- Classroom composition
  - Class size
  - Percent ELL
- Teacher perceptions
  - Student engagement
  - Implementation success
  - Implementation for high achievers

    - Implementation for low achievers
    - Implementation for ELLs
- Interaction with *Seeds/Roots* materials
    - Inquiry-based teachers
    - Percent of time spent on hands-on experiences
    - Percent of time spent on reading
    - Percent of time spent on writing
    - Minutes teaching science
    - Teaching experience
- Additional joint effects
    - Inquiry-based teachers and percent of time spent on hands-on experiences (for current PM unit)
    - Inquiry-based teachers and minutes and science instruction

The results presented in Table 28 are reduced parsimonious models that best captured the impact of teacher characteristics. They examine the impact of teacher self-efficacy as well as the aforementioned teacher characteristics, including classroom average performance. For science content, the results in Table 23 indicate that accounting for teacher self-efficacy, mean class pretest performance, and teacher experience, students in treatment classrooms are expected to outperform students in control classrooms by about 1.1 points (Effect Size = 0.21, $p < .05$). There is suggestive evidence that teacher experience has a positive effect on student science outcomes ($p < .10$). Consistent with previous results, there is a positive relationship between vocabulary knowledge and post science results, and this effect is exacerbated in treatment classrooms. That is, student prevocabulary performance impacts post science content performance. The effect of vocabulary in treatment classrooms compared to control classrooms relates to an effect size of about 0.32. This indicates that students with higher preunit achievement fair relatively better in treatment classrooms than in control classrooms. In other words, the difference in science content posttest performance between a student in a control classroom who is one standard deviation above average on the vocabulary pretest compared to a student in a control classroom who is one standard deviation below average on the vocabulary pretest is about 1.8 points. Meanwhile, in a treatment classroom the same two students would be about 3 points apart. This treatment versus control classroom difference is about 0.32 standard deviations on the science content metric, or about a 0.32 effect size.

Table 28

Effect of Teacher Self-Efficacy

Fixed Effects	Science	*signif*	Vocabulary	*signif*	Reading	*signif*
Intercept (control classroom)	15.08	***	17.68	***	6.99	***
Mean class performance (subject-specific)						
Science	0.12					
Vocabulary			0.22	**		
Reading					0.15	*
Treatment effect (+/-)	1.09	*	1.41	***	-0.07	
Self efficacy (PRE)						
Science	0.01		-0.01		0.00	
Literacy	-0.01		0.00		0.02	
Unit supports standards well			0.97	**		
Years of teaching at current grade	0.08	^				
Pretest performance						
Science	0.42	***	0.07	**	0.03	**
Vocabulary					0.11	***
Control classes	0.21	**	0.47	***		
Treatment	0.14	*	-0.09			
Class mean vocabulary	-0.02		-0.03	**		
Reading	0.41	***	0.36	**	0.37	***

Note. $N = 73$ (teachers)
^$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

The vocabulary results indicate that students in treatment classrooms perform significantly better on the postvocabulary assessment than control students by about 1.4 points ($p < .001$). The effect size is approximately 0.33. There is a significant mean classroom vocabulary effect. This implies that students in classrooms with higher preexisting vocabulary skills perform better on the postvocabulary assessment (i.e., the overall class vocabulary level is higher, facilitating better individual performance on the posttest). Individual prevocabulary performance is related to individual posttest vocabulary performance, but there is no difference in this relationship between treatment and control classrooms. In all classrooms, the higher the prevocabulary levels, the lower the pre-post vocabulary relationship.

There is no treatment effect for reading, but, consistent with vocabulary, there is a mean classroom reading effect. This again implies that students' performance on the

postreading assessment is not only impacted by the individual student's pretest performance but also by the class average pretest performance.

It is important to highlight that for all three outcomes, teacher self-efficacy plays no role in student performance.

Implementation

4a) What factors distinguish between successful and less successful implementation of the treatment (i.e., number of lessons taught, teacher reports of successful implementation, teacher beliefs that the treatment effectively supported state standards)? We have little objective information relating to treatment implementation but rely on teacher reports to provide potential insight. It is important to reiterate that we examined all of the characteristics noted under Question 3b as well as specifically the number of minutes on specific activities such as hands-on activities or writing.[18] Also, we examine whether effects were different in the ELL-only classrooms than those in mixed classrooms (assuming that implementation likely differed in these classrooms). Tables 29 through 31 summarize the model results for science, vocabulary, and reading, respectively.

The results in Table 29 indicate that examining mixed classrooms (classrooms with no or some ELL students) separately from ELL-only classrooms provides consistent results with respect to the *Seeds/Roots* treatment. However, it should be noted that the average treatment effect estimate for ELL-only classrooms is 3.46, with an estimated standard error of 1.86; the small degrees of freedom limit power to detect significant effects. In mixed classrooms, the number of lessons completed is related to performance ($p < .05$). The results in Table 30 are similar in that there is substantively large treatment effect in the ELL classrooms, but again not statistically significant due to the small number of degrees of freedom. In the mixed ELL classrooms, the vocabulary results indicate that if the teacher felt the unit supported the standards well, posttest performance was increased. Also, the number of lessons has a slightly positive relationship with postvocabulary performance in control classrooms and a statistically significantly stronger impact in treatment classrooms (which implies that students learned an additional 0.60 vocabulary words per lesson taught) than with control students.

[18] It is important to note that models including additional teacher characteristics and implementation (etc.) reduce the number of degrees of freedom in the model due to missing data. While treatment and control groups are randomly assigned, this is no guarantee that data are missing randomly. Too much reliance on imputation and/or listwise deletion obfuscates random assignment results.

Table 29

Science Content

	Excluding ELL-only classrooms			9 ELL-only classrooms		
Fixed Effects	Estimate	*se*	*signif*	Estimate	*se*	*signif*
Intercept (control classroom)	15.16	0.33	***	14.09	1.20	***
Treatment effect (+/-)	0.67	0.44		3.46	1.86	
Mean class performance (subject specific)						
Science Content pretest	0.17	0.10				
Vocabulary pretest						
Reading pretest						
Number of lessons	0.05	0.02	*			
Lesson implemented successfully	0.70	0.46				
Pretest performance						
Science Content	0.42	0.03	***	0.46	0.19	***
Vocabulary	0.32	0.04	***	0.02	0.12	
Reading pretest	0.42	0.07	***	0.47	0.08	*

Note. $N = 73$ (teachers)
***p < .001, **p < .01, *p < .05, ^p < .10

Table 30

Vocabulary

	Excluding ELL-only classrooms			9 ELL-only classrooms		
Fixed Effects	Estimate	*se*	*signif*	Estimate	*se*	*signif*
Intercept (control classroom)	17.29	0.37	***	17.49	0.62	***
Treatment effect (+/-)	0.21	0.46		1.79	0.98	
Mean class performance (subject specific)						
Science Content pretest						
Vocabulary pretest	0.20	0.07	**			
Reading pretest						
Unit Support Standards Well	1.10	0.36	**			
Number of lessons	0.04	0.01	**			
Lesson implemented successfully	-0.46	0.35				
Treatment × Lesson implemented	1.49	0.60	*			
Pretest performance						
Science Content	0.06	0.02	**	0.08	0.05	
Vocabulary	0.44	0.03	***	0.37	0.08	***
Reading pretest	0.39	0.06	***	0.25	0.12	*

Note. $N = 73$ (teachers).
^p < .10. *p < .05.**p < .01. ***p < .001.

Table 31 summarizes results for reading. The results indicate, again, that if teachers perceived the unit to support the standards well, then student performance was higher on the reading posttest. There was no such impact in ELL-only classrooms.

Overall, teacher background, training, experience, and method (e.g., inquiry-based) had no systematic impact on student outcomes. The number of lessons as well as the perceived match to state standards tended to impact performance in mixed classrooms but not in ELL-only classrooms. There is some exploratory evidence that for science content and vocabulary, the treatment, on average, tended to have an impact, but statistically these results are inconclusive—primarily due to the small number of ELL-only classrooms.

Table 31

Reading

	Excluding ELL-only classrooms			9 ELL-only classrooms		
Fixed Effects	Estimate	se	*signif*	Estimate	se	*signif*
Intercept (control classroom)	6.79	0.14	***	6.79	0.27	***
Treatment effect (+/-)	-0.03	0.12		0.38	0.45	
Mean class performance (subject-specific)						
Science Content pretest						
Vocabulary pretest						
Reading pretest	0.09	0.06				
Unit supports standards well	0.25	0.14	*			
Number of lessons						
Lesson implemented successfully						
Treatment × Lesson implemented						
Pretest performance						
Science Content	0.03	0.01	*	0.07	0.03	*
Vocabulary	0.13	0.02	***	0.03	0.05	
Reading pretest	0.38	0.03	***	0.32	0.08	***

Note. $N = 73$ (teachers).
^$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

We also examined the impact of classroom composition, in terms of the percent of ELL students in the classroom, and found that, in general, the percent of ELLs in a classroom has no impact on student outcomes except for science content. In no case was the effect different in treatment and control classrooms. Overall, performance in high-percent ELL classrooms (percent ELL ≥ 20%) was lower than in other classrooms. Generally, ELL students perform less well than EO students (all situations).We note that 35% of all students were in classrooms where teachers thought the unit was effective for ELLs. In high-percent ELL classrooms, this percentage is 46% of all students. Table 32 summarizes the impact of percent ELL in a classroom on student outcomes. Again, as noted, only on science content is there a statistically significant effect of classroom composition (percent ELL) and student posttest results.

Table 32

Impact of Percent ELL in Classroom on Student Outcomes

	Effect[a]	*signif*	Effect different in treatment classroom
Science Content	-1.15	*	No
Nature of Science	0.07		No
Inquiry Science	-0.25		No
Vocabulary	-0.17		No
Reading	-0.10		No
Writing	0.10		No

Note. $N = 84$.
[a]The effect compares class mean achievement in a classroom with the percent of ELLs one standard deviation above average against a classroom with the percent of ELLs one standard deviation below average.
^$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

4b) What are teachers' reactions to the quality, usability, and utility of the units? What are some positive aspects of the *Seeds/Roots* Planets and Moon unit; what could be improved? We also examined the impact of the number of lessons among treatment teachers and found that in science and vocabulary (and consistent with results in Table 32) there is some suggestive evidence that the number of lessons plays a role in the impact of the treatment—the treatment being more successful when more lessons are taught. This result is summarized in Table 33.

Table 33

Effect of the Number of Lessons by Treatment Teachers

	Science Content	*signif*	Vocabulary	*signif*
Effect of number of *Seeds/Roots* lessons	0.063	^	0.093	^
Effect size	0.21		0.42	

Note. $N = 38$.
^$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

We surveyed teachers to determine their perceptions of the challenges and potential in teaching PM; additionally, by having the teachers use the *Seeds/Roots* curriculum, we were able to evaluate their reactions to the quality, usability, and utility of the units. Survey responses indicate that all teachers continue to consider making time for science during the school day the primary challenge. However, teachers also noted that keeping students engaged with science and having appropriate materials were also significant challenges. With respect to literacy, teachers were most challenged by the variability in student reading level

and the lack of reading skills in students entering fifth grade. After the PM unit was completed, control teachers remained concerned with time, especially getting the "big" concepts across to the students in the time available. Control teachers continued to have issues with a lack of student vocabulary skills. The biggest challenge for treatment teachers was having enough time to implement the program. With respect to literacy, control teachers continued to have issues with reading, especially vocabulary and comprehension skills. Treatment teachers were most concerned with time and difficulties presented by varying levels of student reading ability. Treatment teachers consistently noted that students were engaged, that there was a "nice mix" of materials, and that students enjoyed opportunities to get out of the book.

In order to more deeply understand teacher perceptions about the *Seeds/Roots* curriculum, a half dozen treatment teachers were interviewed. Interviews were conducted to probe teachers with respect to their experiences teaching the *Seeds/Roots* materials. The teachers interviewed expressed positive reviews of the unit. For example, to the prompt, "Tell me about your experience teaching the unit," teachers in the PM treatment responded that they enjoyed teaching the *Seeds*/*Roots* materials. In addition, the focused literacy component seemed to resonate as value-added. For example, one participant offered, "I really liked the concept development. Liked the emphasis on the language of argumentation. It helped the students buy into the science experiment. The time spent developing concepts helped go in-depth and explore models." In addition, another science teacher stated:

> [I] really enjoyed and liked using the materials. [My] background is hands-on science with a lab everyday [classroom is a science lab]. [The] unit helped ELL get over a hump and was good for students that read at a lower level.

According to the teachers who responded, the students liked the materials. For instance, one teacher responded to the prompt, "Tell me about your student's response to the units," by saying:

> They really seemed to enjoy it. Journaling, concept development, writing a lot, partner pair-share. The students demonstrated an awareness of the moon phases, as they shared in class their observations from the night before. The students seemed more aware of their environment and seemed to apply what they were learning to their own world—though they didn't even get the whole unit.

Another respondent offered: "The students enjoyed it all. Specifically, the hands-on and the reading." However, one teacher who is also a literacy coach suggested that some of the reading materials might have been too difficult for some of the lowest level learners.

When prompted with —What worked well about these materials?" teachers' answers varied with some referring to the content alignment with standards; some referred positively to the organization of the teachers' manual. For example, —Bookswere well designed (Graphics, Charts, Topics). Fit nicely with the standards. Liked the journals." Furthermore, another respondent added, —Likedthe in-depth investigations. The student materials were great." Another respondent offered, —Easyto use. Straight-forward." However, to the follow-up question, —What did you find challenging?" one respondent offered that the length of time necessary to complete the unit was a challenge. Another teacher offered the following: —The journals came apart, which became a challenge. The teacher manual was a challenge—too much reading—list the materials needed. Needs a better format, but good for new teachers." Another teacher echoed the sentiment about the bulk of the materials in the following statement: —Itwas challenging working 30 years. Had a lot of materials to rely on and did not review the materials that accompanied the unit. Would have had a different format."

To the question, —Howeasy were these materials to use?" one teacher suggested, —Tha teachers guide was excellent, bad layout though. The labs should all be on one page. Was heavy on literacy." Another participant —was confused about where to find things" but —was happy with the explanations given for the lessons." Another respondent suggested:

> [The] unit is too long, took more than 40 days to complete. Maximum is 30 days. Not specific enough to (STATE) standards, but too in-depth. Good balance between literacy and traditional inquiry, and reading and research. Shows students that much of what scientists do is read. Has a good structure.

Another teacher noted the opposite: —Easyto understand. The length was manageable. Had more literacy than inquiry."

When asked, —Howdid your use of the unit influence your thinking about teaching science and literacy?" one respondent offered, —Themandatory 3hr literacy block has to integrate science and social science into that framework. Will use literary strategies to assist students with journaling scientific ideas." Another suggested that the unit —had a positive influence, decided to bring to principals attention. Good to use as justification for science during part of literacy." Furthermore, another respondent stated, —Helpedwith ELL. Reading goals will need to be incorporated in coming years. Will use Quick writes and methods to memorize." While some teachers suggested that the units reinforced what they already practiced, another suggested that it —was new to provide books for the curriculum as opposed to just hands-on."

To the question, "As a result of using the unit, has your science instruction changed?" one respondent reported that the writing component was too heavy and that they "will not do so much writing in the future." However, another respondent suggested the opposite: "Method of instruction will change. Will incorporate student science journals. Have reading goals. Quick Writes." Finally, respondents suggested that they would focus more on literacy in the coming semesters. With respect to additional support materials, the teacher responses suggest that additional online support and applications to "Smart Boards" would be welcome. In addition, online links to NASA and local space centers were suggested, along with greater instructional differentiation in the materials.

When prompted with "Having taught this unit, what do you think about integrating science and literacy?" teachers emphasized the strategic alignment needed for the mandatory literacy time commitments in addition to the natural link between literacy training and science. For instance, one teacher offered, "It's a really good idea. Have to link and integrate stuff. Also, the basic needs of students are unmet, lots of emotional problems, would like to emphasize time management more. Reading underlies all aspects." Another said, "Loved it. Believe it is the only way to go with the 3hr block. Delivers more bang for the buck. Has attended meetings and suggesting this as the way to go." Furthermore, another participant offered that "it was a natural fit. Not a stretch. The reading and science integrate." Finally, another teacher said, "The student gets science for reading, allows them to memorize better."

In general, participants' responses were positive. Although not mutually exclusive, participants related the integration of science and literacy to strategic response, to increased reading mandates, as well as to the "naturalness" of the fit between literacy and science instruction. Although not systematically appearing in the quantitative results, implementation fidelity could be linked to the number of years a teacher has taught. At least, based on this small sample, the longer a teacher has taught, the less likely the teacher is to follow the directions. There seemed to be a consensus regarding a review of the teacher's manual, aligning all instructions to one-page documents. In addition, efforts might be useful in designing more rugged "student consumables" (student journals).

Conclusion

The evaluation of the *Seeds/Roots* Planets and Moons curriculum addressed several aspects related to potentially substantively important effects on student performance. Examining several outcomes and several related concomitant factors provided an opportunity to check the robustness of results. Overall, given the random assignment of teachers to the *Seeds/Roots* treatment, a simple test (accounting for clustered design) provides the most

unequivocal answer as to whether or not there was a statistically significant and substantively meaningful effect associated with the treatment. Ideally, all subsequent analyses corroborate the initial findings. In this case, the findings were relatively consistent across the myriad of analyses, with some highlighting potential avenues of a successful intervention, while most, however, consistently pointed to inconsistent impact on outcomes related to the treatment.

The analyses focused on several outcomes of interest, including science content, reading, vocabulary, writing, student attitudes towards science, and teacher self-efficacy. The design is powerful in that teachers were randomly assigned to treatment and control groups, and descriptive analyses indicated that, based on observable characteristics, there were few differences between treatment and control classrooms, preunit. However, the implemented design faced several challenges. First, implementation took place in 10 states—meaning that the comparison was an average business-as-usual rather than a well-defined control group. This, as noted, makes it difficult to pinpoint what elements may have contributed to results (Campbell & Stanley, 1963). Second, postrandomization was a challenge. At one school, control students were placed into treatment classrooms. These classrooms had to be eliminated. Although not a large number (an approximately equal number of control and treatment classrooms were affected), it did reduce the sample size—impacting power to detect effects. Third, the extreme difficulty in obtaining student standardized results and demographic information limited analyses into potential effects of the treatment on specific student subgroups. Qualitatively, this was clearly a result of the prevailing economic conditions most districts and school faced and also of a heighted sense of proprietorship over student information.

The balanced approach to the *Seeds/Roots* unit was examined by analyzing both science and literacy outcomes. The science outcomes consisted of science content, the nature of science, and science inquiry assessments. The science content assessment was the primary science outcome of interest and was examined utilizing two scoring approaches: a traditional multiple choice equally weighted composite and an ordered multiple choice polytomous scoring model that included ten polytomously scored items. The nature of science and science inquiry assessments had fewer items and somewhat lower reliability and served primarily to examine the pattern of results and check robustness of findings. Literacy outcomes included reading, vocabulary, and writing assessments. In general, assessments have sufficient reliabilities to form the basis for analyses. In addition, the evaluation examined the extent to which student attitudes about science were impacted by the *Seeds/Roots* curriculum and whether teacher self-efficacy was affected by the *Seeds/Roots* treatment.

In order to test whether the *Seeds/Roots* treatment had a statistically significant and substantively important effect on student outcomes, a multilevel random effects model was used to account for the fact that the study was a cluster randomized trial—with teachers (clusters) assigned to the treatment or control groups and students attending classes (nested) with(in) teachers. We compared residual performance, that is, posttest outcomes accounting for pretest performance between treatment and control groups. We also utilized the IRT scores from the ordered multiple choice scoring model in order to model true initial status and gain to provide another model for testing treatment effects.

The analyses of science outcomes provided equivocal evidence that the *Seeds/Roots* treatment impacted science outcomes. There was suggestive evidence that science content was affected by the *Seeds/Roots* curriculum ($p < .10$) when examining the research question under the original hypothesis, significant difference in gains ($p < .05$), stronger evidence that the nature of science results were systematically impacted by the *Seeds/Roots* curriculum ($p < .05$), and no effect on science inquiry. The IRT score-based models corroborated the simpler multiple choice results, to a large extent. The IRT-based gain models implied that student performance decreased slightly, about -0.06 ($p < .01$), in control classrooms but that students in treatment classrooms performed slightly better ($p < .10$) by staying about the same (-0.02) from the pretest to the posttest. Assuming the polytomous scoring models adequately reflect connections among concepts and that the scores provide results from which we can validly infer conceptual understanding, the results imply no true gains when including indicators of conceptual understanding. The fact that the multiple choice and ordered multiple choice models result in a different overall picture of student performance (the former indicating improvement with the latter indicating no improvement) is potentially a result of the differential impact of basic factual recall and conceptual understanding. A few teachers noted that the *Seeds/Roots* reading materials provided additional support to help students "memorize" important elements of the unit. Without independent observations of control and treatment classrooms, we can only speculate the extent to which teachers emphasized conceptual understanding of materials or memorization (in either condition).

More complex models accounting for concomitant factors as well other aspects of implementation further corroborated results in that student performance (in at least vocabulary) was improved if the teacher felt that the unit supported the state standards well. This was consistent across treatment and control classrooms. Further, disaggregation of results (into classrooms with both EO and ELL students and classrooms with only ELL students) indicated that in the mixed classrooms there is some evidence that if the treatment teachers (but not control teachers) felt the lesson was implemented successfully, then this

was associated with improved performance (however, treatment teachers were much less likely to perceive that their lessons were implemented successfully). Overall, the number of lessons was also positively related to student outcomes in science content and vocabulary—irrespective of teacher background. Specifically, among treatment teachers there was also suggestive evidence that the number of lessons was related to science content and vocabulary outcomes ($p < .10$).

Consistent with much prior research on teacher effects, there was little relationship among reported teacher background, credentialing and experience, and student outcomes. Although previous research indicates that teacher content knowledge in mathematics and reading is positively related to student performance, the self-reported measures were not related to the science and literacy outcomes.

We also evaluated the impact of the *Seeds/Roots* curriculum on literacy. The *Seeds/Roots* curriculum had an unequivocally positive effect on student vocabulary ($p < .01$). The effect size of approximately 0.40 demonstrates that this effect was substantively meaningful. Overall, there was no treatment effect on either reading or writing. One contributing factor likely related to overall reading and writing performance could be both the variability and lack of reading and vocabulary skills possessed by students. Empirical results indicate that preexisting vocabulary played an important moderating role in other student outcomes. Moreover, preexisting vocabulary played a greater role in treatment classrooms than in control classrooms. The *Seeds/Roots* curriculum had a greater impact with students on science content posttest results when they had stronger preexisting vocabulary skills, which is consistent with the emphasis on integrating reading and writing in the *Seeds/Roots* curriculum. Still, for all students, vocabulary gains were related to better posttest science content and reading performance. Although this effect was equal in treatment and control classrooms, given that there was more vocabulary improvement in treatment classrooms, we would expect this as another mechanism through which the *Seeds/Roots* curriculum might impact science and literacy.

However, preexisting skills are a hurdle that impacts all teachers. As noted, variability and lack of literacy preparation was a challenge that was consistently voiced by treatment and control teachers alike. Teacher perceptions that the *Seeds/Roots* unit was better suited for high achievers were corroborated by the evidence (particularly as measured by vocabulary). Moreover, about 50% of teachers in both conditions indicated that the challenge level of the unit was inappropriate for students. Of equal concern is that a few teachers noted that the reading materials helped students "memorize" facts. This may not have been what the

treatment teachers meant explicitly, but, as noted, this emphasis certainly provides some explanation as to the lack of consistent treatment effects.

The evaluation also attempted to examine the impact of the *Seeds/Roots* curriculum on specific student subgroups. As noted previously, we were constrained in our ability to thoroughly conduct these analyses, but were able to examine the impact of the *Seeds/Roots* curriculum on ELL students—both directly and indirectly. The results using individual student identifiers that classified them into ELL and EO students revealed that, on average, ELL students faired about as well as their EO classmates. This is not to say that there were no performance gaps, but treatment had roughly equal impact on EO and ELL student science and literacy performance. The *Seeds/Roots* treatment did not close any existing performance gaps.

We also took advantage of the fact that a small subset of teachers reported having classrooms consisting of 100% ELL students. We reexamined the outcomes splitting the analyses by classroom type (100%, or ELL-only Classrooms, and mixed ELL/EO classrooms). The treatment effect on science content was, on average, very large in ELL classrooms but was likely not significant due to the small degrees of freedom (estimate 3.46, standard error 1.86). It is important to note that previously discussed results related to implementation remain significant in the mixed classrooms but not in the ELL-only classrooms. Also, the vocabulary transfer effect on science content was substantively smaller and nonsignificant in the ELL-only classrooms. While the *Seeds/Roots* curriculum was not specifically designed for ELL-only classrooms, the suggestive evidence indicates that the balanced science and literacy approach potentially provides positive support for ELL students. This clearly warrants additional research.

We examined nonacademic outcomes for both students and teachers and found that there was little change in either student attitudes towards science or teacher self-efficacy for teaching science and literacy. Neither was impacted differentially by the *Seeds/Roots* curriculum.

Overall, the results for the PM unit were not as unequivocal as previous *Seeds/Roots* interventions (Goldschmidt & Jung, 2009; Wang & Herman, 2006), but this may also have been partially due to the fact that the study occurred in ten states. Although the study was not designed to examine state effects, exploratory analyses revealed that teachers within states did not uniformly respond to survey questions, which provides some insight into the nature and potential impact of teacher perceptions. The state results also highlighted that the intervention worked differentially well in different states. The reasons for these differences

likely go beyond simply how well the unit supports state standards, but again demonstrate the importance of considering context in developing and implementing treatments in different settings.

Consistent with previous *Seeds/Roots* evaluations were teacher responses related to time to complete the unit. Teachers reported spending about 50% more time on the *Seeds/Roots* materials than on the state standard unit, which is consistent with the integrated notion of the treatment. However, also consistent with previous evaluations is that teachers overwhelmingly would use the unit again (some with modifications and time constrains) and felt the materials were very good, helped engage students, and reduced, to a large extent, a concern voiced by control teachers related to the lack of materials available for science instruction. In contrast with previous results (Goldschmidt & Jung, 2009), only about half of the teachers felt comfortable with the *Seeds/Roots* materials, whereas previously about two-thirds of teachers felt comfortable with the *Seeds/Roots* materials.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1–43.

Briggs, D., Alonzo, A., Schwab, C. & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33–64.

Bryk, A. S., Thum, Y. M., Easton, J. Q., & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education*, *2*, 103–142.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In *Review of Research in Education* (Vol. 8, pp. 158–233). Washington, DC: American Educational Research Association.

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi experimental designs for research*. Chicago: Rand McNally.

Cervetti, G. , Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. (2009, April). *Integrating science and literacy: A value proposition?* Paper presentation at the annual meeting of the American Educational Research Association, San Diego, CA.

Cervetti, G., Pearson, P. D., Bravo, M. A., & Barber, J. (2006). Reading and writing in the service of inquiry-based science. In R. Douglas, M. Klentschy, and K. Worth (Eds.), *Linking Science and Literacy in the K-8 Classroom*. Arlington, Virginia: National Science Teachers Association.

Goldschmidt, P., & Jung, H. (2009, June). *Evaluation of Seeds of Science/Roots of Reading: Effective Tools for Developing Literacy through Science in the Early Grades, Final Interim DRAFT Deliverable.* National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), Graduate School of Education & Information Studies, University of California, Los Angeles.

Guthrie, J. T., & Ozgungor, S. (2002). Instructional contexts for reading engagement. In C. C. Block & M. Pressley (Eds.), *Comprehension Instruction: Research-based Best Practices.* New York: Guilford Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Romance, N. R., & Vitale, M. R. (1992). A curriculum strategy that expands time for in-depth elementary science instruction by using science-based reading strategies: Effects of a year-long study in grade four. *Journal of Research in Science Teaching*, *29*(6), 545–554.

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The Handbook of Quantitative Methods for the Social Sciences* (pp. 259–280). Thousand Oaks, CA: Sage Publications.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Stoddart, T., Pinal, A., Latzke, M., & Canaday, D. (2002). Integrating inquiry science and language development for English language learners. *Journal of Research in Science Teaching*, *39*(8), 664–687.

Wang, J., & Herman, J. (2006). *Evaluation of Seeds of Science/Roots of Reading Project: Shoreline Science and Terrarium Investigations.* CSE Technical Report 676, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), Graduate School of Education & Information Studies, University of California, Los Angeles.